DOCUMENT RESUME

ED 202 863 TM 810 279

AUTHOR Matthews, Tom A.; And Others
TITLE Use of Districtwide Test Scores to Compare the Academic Effectiveness of Schools.
INSTITUTION Seattle Public Schools, Wash. Dept. of Planning, Research, and Evaluation.
PUB DATE Apr 81
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Educational Quality; Elementary Education; Evaluation Criteria; Institutional Characteristics; *Multiple Regression Analysis; *Program Effectiveness; *School Districts; *Scores
IDENTIFIERS *Index of Academic Gains

ABSTRACT

The Evaluation Office of the Seattle Public Schools is currently engaged in a project that uses student test performance to identify educational program characteristics associated with academic gains. Methods used to develop an index of academic gains are described, and one method that best fits the following criteria was selected. It should be: (1) a measure of student gain in academic subjects such as reading and math in a single school year; (2) applicable to schools (rather than specific educational programs) as the primary educational unit; (3) relatively free of the effects of factors beyond the control of schools; (4) based upon sound methodological procedures; (5) adaptable to the test data currently on hand; and (6) easy to interpret and explain. The method finally selected was multiple regression of spring score on fall score, ethnic category, and free lunch eligibility, with school mean residual as the measure of academic gain. Steps currently being implemented to validate the index in school settings are discussed. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED202863

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

USE OF DISTRICTWIDE TEST SCORES TO COMPARE THE ACADEMIC EFFECTIVENESS OF SCHOOLS

Tom A. Matthews, Jane B. Soder,
Madelaine C. Ramey, George H. Sanders,

Seattle Public Schools

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

T. Matthews

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of
the American Educational Research Association

April 1981

TM 810 279

USE OF DISTRICTWIDE TEST SCORES TO COMPARE THE ACADEMIC EFFECTIVENESS OF SCHOOLS

The Evaluation Office of the Seattle Public Schools is currently engaged in a project that uses student test performance to identify educational program characteristics associated with academic gains. Such information is desirable to make and defend decisions about which aspects of educational programs should be continued, expanded and/or transported to other sites, which should be studied more intensively, and which should be modified or discontinued. Of particular interest for decision making are educational program characteristics that are manipulable; i.e., are under District and/or school control.

The largest portion of this report describes the methods used to develop an index of academic gains. Further, the paper discusses steps currently being implemented to validate the index in school settings. Possible uses or modifications of the index will be dictated, in large part, by the results of the validation study. If the index is strongly related to school level characteristics which have been shown in the literature to produce large academic gains, it can be used with confidence to identify exceptionally effective schools. However, if the index scores are not related to previously identifed characteristics of effective schools, further modifications of the index or methods used to assess school characteristics are warranted before the index can be used to provide information of practical consequence.

Various authors have proposed alternative models of school achievement (Harnischfeger and Wiley, 1977; Centra and Potter, 1980). Typically these

models view student achievement as a function of some combination of school and nonschool effects; i.e., school size, resources, facilities, teacher characteristics, instructional organization, pupil-teacher ratios, student characteristics, etc.. The present effort builds upon the distinction between school and nonschool effects in an effort to develop a useful index of academic gains. Basically, an attempt has been made to statistically control variance associated with major nonschool characteristics (Socio-Economic Status or SES, initial achievement, and ethnicity), thereby creating an index or residual that is independent of the combined effects of major nonschool variables. Although the authors recognize that the resulting index also reflects some of the effects of uncontrolled nonschool variables, all possible interactions, and random error, and that the index has been purged of some school effects that are related to major nonschool characteristics, the index is treated as a reasonable, initial indicator of academic gains for schools.

DEVELOPING A STATISTICAL MEASURE OF ACADEMIC GAIN

The process used to develop the statistical measure (index) of academic gain warrants description. First, the criteria that an acceptable index should meet were identified. Then, a number of alternative methods of measuring academic gain were investigated and a method that best fit the criteria was selected. The criteria developed for the statistical measure of academic gain are these:

1. It should be a measure of student gain in academic subjects such as reading and math in a single school year.
2. It should be applicable to schools (rather than specific educational programs) as the primary educational unit.
3. It should be relatively free of the effects of factors beyond the control of schools.
4. It should be based upon sound methodological procedures.

5. It should be adaptable to the test data currently on hand.

6. It should be easy to interpret and explain.

The literature concerning each of the commonly used methods for assessing school effects was reviewed and critiqued by the authors on methodological and theoretical grounds. Some methods were based on statistical techniques widely used in program evaluation and research such as raw score gain, standardized change, and estimated true gain. Some were methods that have been used by school districts or research organizations to identify effective schools. These included widely used methods such as covariance and regression (Sanders and Crockett, 1976; Klitgaard and Hall, 1973; Jencks and Brown, 1975; Dyer, Linn, and Patton, 1969; Rutter, Maughan, Mortimore, and Ouston, 1979), and less well known techniques such as the equipercentile method (Soder, 1979) and "Idiot's Regression" (Dade County Public Schools, 1972; Stephenson, 1973). Additional variations were also developed by the Seattle Public Schools' Evaluation Office staff.

All of the methods were based on fall and spring achievement test scores in a single year. The focus was either on test score change during the year or deviation from the spring score expected on the basis of fall score and sometimes other variables such as SES, sex, or ethnic category.

Developing a Procedure

Each of the methods was applied to 1978-79 grade 6 Total Reading scores on the Metropolitan Achievement Test. Six low and six high "outlier schools" were identified using each of the methods. These were the schools with the lowest or highest gain or the largest deviations from expected score, depending on the method.

There was considerable overlap among the schools identified by different methods but in no case did two methods select exactly the same schools. In two cases, a school was identified as a high outlier by some methods and as a low outlier by others. This finding indicated that caution should be used in interpreting the results of the method finally selected.

Each of the methods yielded a measure or index score for each school. The measures were change scores for some methods and deviation or residual scores for others. Correlations between scores resulting from each pair of methods were computed to get a statistical indication of the degree of similarity of results among the methods. These correlations suggested that there were three groups or "clusters" of methods, with each cluster containing several methods. The index scores produced by methods within each cluster correlated very highly (.93 to .99) with each other. Index scores of methods in different clusters were moderately correlated (.4 to .5).

One of the criteria for the index of academic gain was that it should not be affected by factors outside the control of schools. Correlations were computed between the school-by-school index scores produced by each method (using the grade 6 MAT Total Reading scores) and mean fall reading scores, percent White, and percent low SES (using free lunch eligibility as an indicator). None of the methods produced school index scores that were correlated very highly with mean fall reading scores. Correlations ranged from -.18 to +.19, and none was statistically significant at the .05 level (37 schools were used in the analysis).

Likewise, no method produced scores that had a statistically significant correlation with SES as measured by free lunch eligibility. Most correlations were in the direction of larger gain associated with smaller percentages of students eligible for free lunch. The most extreme correlation was -.22.

Two methods produced scores significantly correlated with percent White ($.30<r<.34$). All other correlations with percent White ranged between $\pm.22$ and were nonsignificant at the .05 level (n=37).

The method finally selected for use in providing a statistical measure of academic gain was multiple regression of spring score on fall score, ethnic category, and free lunch eligibility, with the school mean residual as the measure of academic gain. This method was applied separately to test data from each grade level between two and six and further refined. The two major refinements were the addition of polynomial terms as independent variables and the application of a test of significance to the mean residual scores.

A trend analysis revealed a small deviation from linearity when using multiple linear regression. An analysis of deviation scores by fall score revealed that deviations tended to be slightly negative for students initially scoring high or low, and slightly positive for students scoring in the middle of the distribution. This would result in some bias against schools with large numbers of high-scoring or low-scoring students. Therefore the decision was made to include polynomial functions of fall test scores in the regression equation. For each grade level and subject area, all (up to 5) significant polynomial terms were included. With the addition of these terms, the mean deviation from expected is near zero at all points on the fall score distribution.

If school mean residuals were used as index scores, one could not tell whether a relatively large negative or positive score was a true indicator of large or small academic gain, or whether this score was produced largely by measurement or sampling error. Neither would these scores provide a control for school size. Thus, the next step was to apply a test of significance to each school's mean residual score (within each grade level) in order to identify high and low

"outliers"; i.e., those schools scoring significantly above or below expectation.* The test consisted of dividing each school's mean residual score by the estimated standard error of that score and using the standardized normal distribution to assess the significance of the resulting statistic.** Those schools with absolute values of the statistic exceeding 2.58 (=.01) were considered high or low outliers. Since the test statistic, being in standard score units, is comparable from school to school and grade level to grade level, it was decided to use it rather than the mean residual score as a school's grade level index of academic gain.

---

* It might be argued that, since schools constitute an exhaustive partition of an entire population (all students tested within a grade), there is no generalization to be made and so a significance test is superfluous. However, as Blalock (1972) suggests, it is conceivable that some other partition, even one produced by a random process, could give rise to differences as large or larger than those obtained. Thus the purpose of the significance test was to rule out this "simple 'chance-process' alternative." (pp. 238-239)

** The standard error of the mean residual score was estimated by

$$\sqrt{\frac{MS_{RES}}{n}(1-n/N)}$$

where MS_{RES} is the residual mean square from the multiple regression analysis, N is the population size (all students tested within a grade) and n is the sample size (the number of students tested within a grade in a given school). Under the null hypothesis of no school-specific effect and the assumption that residual scores are distributed approximately normally, the mean residual score divided by the above estimated standard error is distributed approximately as a standardized normal variate. (Because n was very small relative to N, the term (1-n/N) was dropped in the actual calculations.)

The resulting index appeared to be consistent with the the following criteria established for selecting an index of academic gains.

1. It can be used as a measure of student gain in academic areas such as reading and math in a single school year. By using spring reading or math scores as the dependent variable and fall scores as an independent variable, the residual score can be interpreted (in part) as a measure of gain.

2. It can be applied to schools as the unit of analysis. The mean residual score for each school can be used to produce an index score. (For example, see Jencks and Brown, 1975 and Klitgaard and Hall, 1973).

3. It appears to be relatively unaffected by factors beyond the control of schools. Certain nonschool factors (the beginning achievement level, SES, and ethnic category of students) are directly controlled by making them independent variables. The method produces mean residual scores of near zero for every ethnic category, every level of beginning achievement, and for students eligible for and not eligible for free and reduced-price lunch.

4. It is based upon well established statistical procedures. The literature review indicated less serious problems with the regression approach than with methods based upon raw gains, (Cronbach and Furby, 1970). The possibility of its producing artifically low index scores for initially low-scoring schools (Campbell and Erlebacher, 1970) was partially ruled out by the small, non-significant correlation of index scores and fall test scores and the near-zero mean residual for groups of students at all fall score levels.

5. It can be easily computed using scores from achievement tests given districtwide in the fall and spring. These scores, as well as free lunch eligibility and ethnic category are all readily available for computer analysis.

6. It is relatively easy to interpret and explain. Because it is a well known statistical technique and because it can be depicted graphically with good effect, multiple regression and the use of school mean residuals can be explained to educators without strong statistical backgrounds.

APPLICATION OF THE INDEX OF ACADEMIC GAIN

The index was applied to test score data from the 1978-79 and 1979-80 school years. For 1978-79, Total Reading and Total Math subtest raw scores on the Metropolitan Achievement Test for grades 2, 4, 6, and 8 were used; for 1979-80, the measure was applied to Total Reading and Total Math subtest raw scores on the California Achievement Test for grades 2-8. Index scores for each school were computed for each grade and subject area.

Figures 1 (reading) and 2 (math) show the distribution of 1979-80 reading and math index (Z) scores for elementary programs in grades two through six. In both subject matter areas, extreme index scores suggest the existence of unusual patterns of academic gain which deserve further investigation.

To investigate the consistency of results from applying the index, five schools with the highest and five schools with the lowest index scores at each grade level and subject area were identified for each of the two school years. No schools were consistently high or low in each subject area by grade level com-

FIGURE 1

FREQUENCY DISTRIBUTION OF 1979-80 READING Z SCORES
FOR EACH TESTED GRADE LEVEL--IN EACH
ELEMENTARY PROGRAM (N=327)

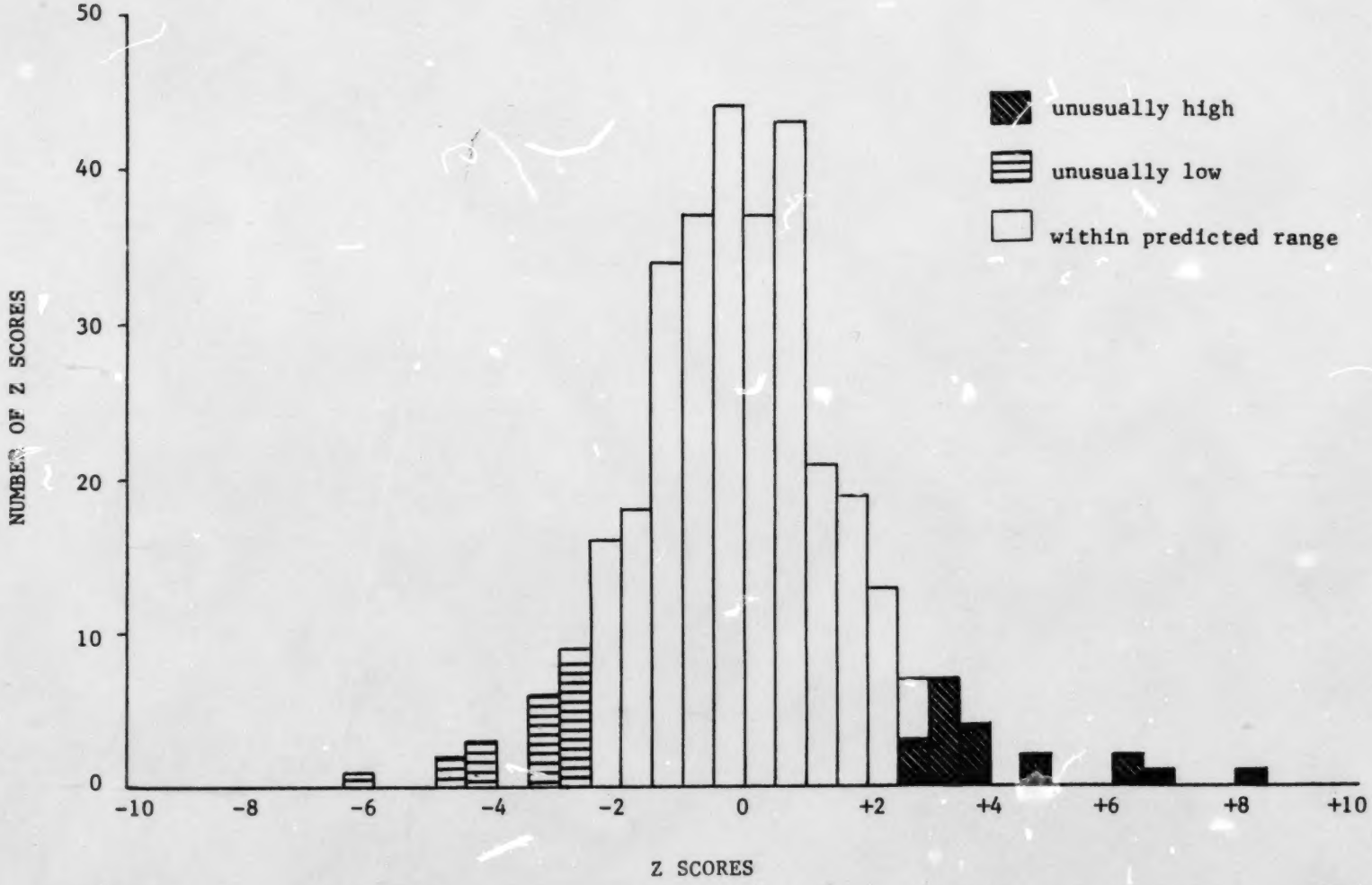
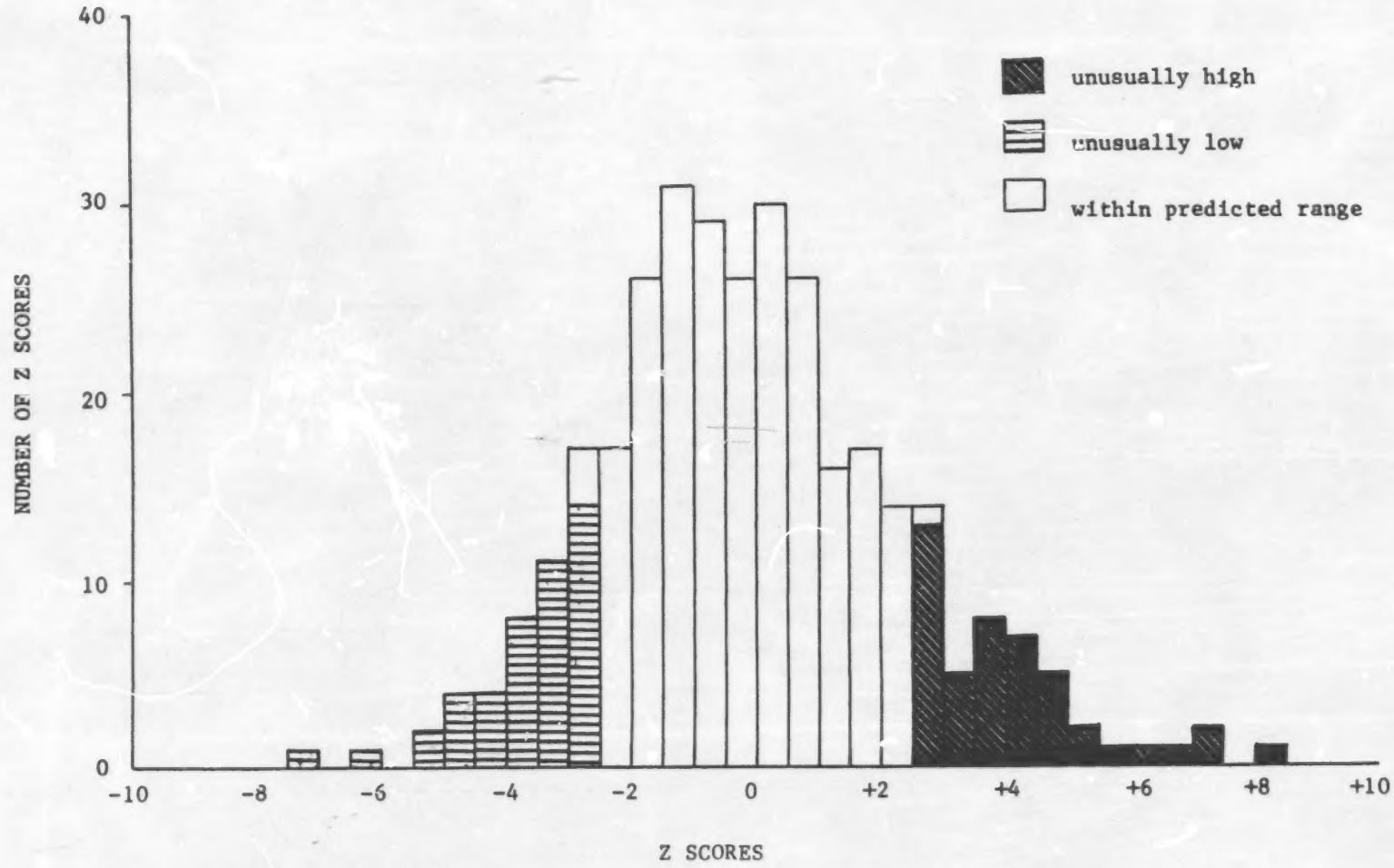


FIGURE 2

FREQUENCY DISTRIBUTION OF 1979-80 MATH Z SCORES
FOR EACH TESTED GRADE LEVEL--IN EACH
ELEMENTARY PROGRAM (N=327)



bination (unless only one grade level was tested). Some were high or low in both subjects at one grade level however, or in reading or math in two or more grades.

Some schools were high or low in both school years. Of these, some schools had extreme index scores in different grades or subjects for the two school years. In some cases, schools were high in one year and low in the other.

Because each school had an index score for each year, it was possible to correlate the 1978-79 scores with the 1979-80 scores for grades 2, 4, 6, and 8. These correlations ranged from -.24 for grade 8 math to +.44 for grade 2 reading. The correlations for both reading and math in grade 8 were negative. These correlations were small and not statistically significant (n=18). For grades 2, 4, and 6, the correlations were positive, but only low to moderate in magnitude.

Correlations between index scores for two school years pair different student cohorts, so a high correlation would point toward school effects. The low correlations obtained here indicate that high or low performance at a given grade level in a school may have more to do with characteristics of that particular student cohort than with school effects. On the other hand, if a school did turn out to be a high or low outlier for both years, this would be evidence for the importance of school effects.

The 1979-80 grade level index scores were averaged to obtain overall reading and math scores for each school. To insure that schools with very high or very low averaged grade level index scores were statistically significant outliers, a statistical test was (again) applied. The test was based on the fact that the average of mutually independent standardized normal variates divided by the

square root of the number of terms averaged is also a standardized normal variate.

Since, as was noted earlier, the grade level index scores are distributed approximately as standardized normal variates and since these index scores were determined independently for each grade, the average of the grade level index scores for a given school divided by the square root of the number of grades tested is distributed approximately as a standardized normal variate. Thus the standardized normal distribution was used to assess the significance of the overall school index scores. The lists of high and low outlier schools using these scores included many, but certainly not all, of the schools which had high or low grade level index scores.

PROPOSED VALIDATION ACTIVITIES

A statistical method for assessing school level academic gains has been proposed. The next steps are to collect data on pertinent school variables and explore the relationships between these variables and the index of academic gain. This process will facilitate the identification of school level variables that affect academic gain and the establishment of the validity of the index.

Because the index could play an important role in making decisions about program improvement, it is essential that its value and educational meaning be established. While it is known that the index is not related to student ethnic category, economic status, or prior achievement level, it is not known to what factors it is related. It is expected that differences in educational programs contribute to variations in the index. It is also possible that test administration procedures or programs designed for special student populations lead to some of the results; or, that student differences not accounted for by the

statistical procedure, such as physical handicaps or emotional problems, would lead to unusual index scores. The following activities are being completed during the 1980-81 school year to refine the index and determine the conditions under which it can be most effectively used:

1. Identifing pertinent school characteristics that can be directly observed. Develop observation techniques.
2. Training observers to use observation techniques.
3. Observing and measuring school characteristics.
4. Examining the relationship between observation results and the statistical index.
5. Reporting on the relationships between school characteristics and the index and how these relationships clarify the meaning and value of the index.

It is expected that some educational program characteristics will be related to the statistical index if it is a valid measure of effectiveness. The following group of characteristics will require direct observation to measure:

1. Principal and teacher emphasis on basic skills.
2. "Time on Task" in reading and math.
3. Instructional organization and structure.
4. Student motivational strategies.
5. Teacher expectations for students.
6. Leadership style of principal.

Other factors that may be related to index scores can be measured centrally. These include:

1. Staff turnover.
2. Ethnic distribution.
3. Student:teacher and teacher:administrator ratios.

Input concerning pertinent characteristics should come from a wide range of sources including school administrators, teachers, parents, and previous research. Standard measurement techniques must then be developed so that program observers can accurately and consistently assess the selected characteristics.

In summary, this paper describes the methods used to develop an index of academic gain that can be applied at the school level. The index is designed to statistically control for initial student differences in achievement levels, SES, and ethnicity using multiple regression techniques. Further, methods used to aggregate data to the school level of analysis are reported.

Because the authors feels that the index should be validated, procedures are described for ongoing validation activities.

REFERENCES

Blalock, H.M., Social Statistics. New York: McGraw Hill, 1972.

Centra, John A. and David A. Potter, "School and teacher effects: An interrelational model." Review of Educational Research, Summer 1980, Vol. 50, No. 2.

Cronbach, Lee J. and Lita Furby, "How should we measure 'change' - or should we?" Psychologial Bulletin, 1970, Vol. 74, No. 1.

Dade County Public Schools, Evaluation Section, Achievement in Dade County Public Schools, 1971-72. Miami, Florida: November 1972.

Dyer, Henry S., Robert L. Linn, and Michael J. Patton, "A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests." American Educational Research Journal, Vol. 6, No. 4, November 1969.

Harnischfeger, A. and Wiley, D.E. "Conceptual issues in models of school learning." Studies of Educative Processes, No. 10. Chicago, Ill.: CEMREL, Inc., 1977.

Jencks, Christopher S. and Marsha D. Brown, "Effects of high schools on their students." Harvard Educational Review, Vol. 45, No. 3, August 1975.

Klitgaard, Robert E. and George R. Hall, A Statistical Search for Unusually Effective Schools. Rand, Santa Monica, California: March 1973.

Rutter, Michael, Barbara Maughan, Peter Mortimore, Janet Ouston, Fifteen Thousand Hours. Secondary Schools and Their Effects on Children. Harvard University Press, Cambridge, Massachusetts: 1979.

Sanders, George and Geraldine T. Crockett, Expected Gains on the Metropolitan Achievement Test. Report No. 76-22, Seattle Public Schools, November 1976.

Soder, Jane B., Expected and Achieved Spring Test Scores by School, 1978-79. Report No. 79-22, Seattle Public Schools: December 1979.

Stephenson, Bob, Expectancy Tables and Pupil Achievement: A Method of Adjusting for Past Performances, Dade County Public Schools, 1973.